The Thesis committee for Whitney Gail Smith

Certifies that this is the approved version of the following thesis:

The C-terminal DNA endonuclease region and biotechnology applications of a group II intron reverse transcriptase from *Thermosynechoccus elongatus*

APPROVED BY

SUPERVISING COMMITTEE:

Supervisor: _____
Alan Lambowitz

_____
Scott Stevens

The C-terminal DNA endonuclease region and biotechnology applications of a group II intron reverse transcriptase from *Thermosynechoccus elongatus*

by

Whitney Gail Smith, B.S.

Thesis

Presented to the Faculty of the Graduate School

of the University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the degree of

Master of Arts

The University of Texas at Austin

August 2011

The C-terminal DNA endonuclease region and biotechnology applications of a group II intron reverse transcriptase from *Thermosynechoccus elongatus*

by

Whitney Gail Smith, MA

The University of Texas at Austin, 2011

Supervisor: Alan M. Lambowitz

Group II introns insert site-specifically into DNA target sites through a process termed retrohoming. They consist of a structured, catalytically active intron RNA and its encoded protein. The protein contains several domains, including a reverse transcriptase domain and a DNA endonuclease domain used for bottom-strand cleavage. Recently, the thermophile *Thermosynechococcus elongatus* BP-1 was found to contain eight functional group II intron-encoded proteins. The proteins are thermostable and active at temperatures up to 65°C. The intron-encoded protein, TeI4c displays the greatest reverse transcriptase activity of these eight proteins, as well as high fidelity and processivity; ideal qualities for a commercial reverse transcriptase. This work explores the possibility of using TeI4c for biotechnology applications, and specifically examines the C-terminal endonuclease domain of TeI4c and its effect on reverse transcription. Additionally, this work investigates the retrohoming activity of a TeI4c truncation that deletes the endonuclease domain.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Group II introns are retrotransposons that insert site-specifically into DNA target sites through a process termed retrohoming (Lambowitz and Zimmerly, 2004). They consist of a highly structured catalytic RNA ("ribozyme") and an associated, multifunctional intron-encoded protein (IEP) (Lambowitz and Zimmerly, 2010). Retrohoming occurs through an excised lariat intron RNA reverse splicing into a DNA target site, followed by reverse transcription of the integrated intron into the genome (Figure 1.1) (Smith et al., 2005). Group II introns have been observed to retrohome into specific DNA target sites at frequencies approaching 100%. However, in a process termed retrotransposition group II introns insert into ectopic sites that resemble the DNA target site (Lambowitz and Zimmerly, 2004).

Group II introns are found within the genomes of bacteria and eukaryotic organelles. Under physiological conditions, group II intron splicing is assisted by the intron-encoded protein (IEP). The IEP consists of a reverse transcriptase domain, a maturase domain used in intron splicing, a DNA binding domain, and an endonuclease domain used during second strand cleavage (Figure 1.5) (Lambowitz and Zimmerly, 2010). The reverse transcriptase domain contains eight conserved sequence motifs, RT0 – RT7 (Zimmerly et al., 2001). The sequence motifs RT1- RT7 are also found in retroviral and other RTs. The upstream sequence motif, RT0, is characteristic of the RTs encoded by non-LTR retrotransposons (Malik et al., 1999; Zimmerly et al., 2001).

Intron-encoded proteins are divided into nine major lineages termed mitochondrial, chloroplast-like 1 and 2, and bacterial A-F classes (Simon et al., 2009;

Toro et al., 2002; Zimmerly et al., 2001) (Figure 1.2). The best-characterized group II intron is L1.LtrB from *Lactococcus lactis*, which encodes the protein LtrA. The L1.LtrB intron RNA belongs to structural subclass IIA1 of the mitochondrial lineage (Zimmerly et al., 2001). While LtrA contains an endonuclease domain, not all group II intron-encoded proteins contain the domain (San Filippo and Lambowitz, 2002). Half of the bacterial classes (including all members of C, D, and E), and some of the organelle IEPs do not contain an endonuclease domain, yet many of these introns retain their ability to retrohome (Dai and Zimmerly, 2002a; Martinez-Abarca and Toro, 2000).

1.1 Mechanism of group II intron mobility

Group II introns splice through two sequential transesterification reactions that results in an excised intron lariat with a 2'-5' phosphodiester bond and ligated exons (Peebles et al., 1986; Schmelzer and Schweyen, 1986) (Figure 1.3). The splicing reaction is catalyzed by the intron RNA, which folds into a conserved secondary structure consisting of six double-helical domains (DI-DVI) arranged around one central circle (Michel and Ferat, 1995; Qin and Pyle, 1998) (Figure 1.4). DV interacts with DI to form the active site in the catalytic core. DIV contains the open reading frame that encodes the group II intron protein. DI contains sequences EBS1 and EBS2 (exon-binding sites 1 and 2) which base pair with the 5'- exon sequences, IBS1 and IBS2 (intron-binding sites 1 and 2). The sequence adjacent to EBS1, δ, base pairs with the first nucleotides of the 3'-exon, δ' (Lambowitz and Zimmerly, 2004). These base pairing interactions help position the intron for RNA splicing and reverse splicing (Costa et al., 2000).

Group II intron mobility occurs through several steps involving the intron RNA and IEP (Matsuura et al., 2001; Saldanha et al., 1999). First, the IEP assists intron splicing by stabilizing the catalytically active RNA structure (Figure 1.1). The IEP then remains associated with the excised intron lariat RNA forming a ribonucleoprotein particle (RNP) (Matsuura et al., 2001; Saldanha et al., 1999). The RNP next recognizes the DNA target site by the IEP interacting with the 5'- and 3'- exons target sequences, and EBS-IBS base-pairing interactions with the intron RNA (Mohr et al., 2000; Singh and Lambowitz, 2001). Following recognition of the DNA target site, the intron RNA reverse splices directly into the DNA strand and bottom-strand cleavage is catalyzed by the endonuclease domain (Guo et al., 1997; Mohr et al., 2000). Second strand cleavage creates a primer, which is used to reverse transcribe the intron into single-stranded complementary DNA (cDNA) (Matsuura et al., 1997; Mohr et al., 2000). Lastly, the cDNA is integrated into the genome through cellular DNA recombination or repair mechanisms (Smith et al., 2005).

1.2 Endonuclease-independent retrohoming

The endonuclease domain contains amino acid residue sequence motifs characteristic of the H-N-H family of DNA endonucleases, with one catalytically essential Mg^{2+} ion at the H-N-H active site (San Filippo and Lambowitz, 2002). Additionally, the endonuclease domain contains two pairs of conserved cysteine pairs (denoted $CX_2C/1$ and $CX_2C/2$) that help maintain the domain's higher-order structure (Gorbalenya, 1994; San Filippo and Lambowitz, 2002; Shub et al., 1994). Endonuclease activity is not required for reverse transcription, reverse splicing, or top strand cleavage

(Matsuura et al., 1997; Zimmerly et al., 1995). However, deletion of the conserved endonuclease domain of LtrA inhibits reverse transcription, possibly because it interacts with and stabilizes the reverse transcriptase domain (San Filippo and Lambowitz, 2002). For all characterized group II introns, interaction between the IEP and specific residues in the 3'-exon is required for endonuclease cleavage (Singh and Lambowitz, 2001). The cleavage site for group II introns varies, with LtrA cleaving between 3' exon positions +9 and +10 (Matsuura et al., 1997).

Several IEPs lack the DNA endonuclease domain (Dai and Zimmerly, 2002a; Martinez-Abarca and Toro, 2000). The best characterized is the catalytically active group II intron from *Sinorhizobium meliloti*, RmInt1, which despite lacking the endonuclease domain retrohomes efficiently using endonuclease-independent retrohoming (Munoz-Adelantado et al., 2003). Endonuclease-independent retrohoming occurs by the group II intron reverse splicing into double-stranded or transiently single-stranded DNA followed by reverse transcription of the intron using a nascent DNA strand at the replication fork as a primer (Figure 1.6). The primer can be either a leading (LEAD) or lagging (LAG) DNA strand (Zhong and Lambowitz, 2003). LtrA predominantly retrohomes using endonuclease-dependent bottom-strand cleavage (Dickson et al., 2001). However, LtrA retrohomes at lower frequencies using endonuclease-independent mobility with a leading-strand orientation bias (Dickson et al., 2001; Zhong and Lambowitz, 2003).

1.3 Group II introns from *Thermosynechoccus elongatus*

The thermophilic bacteria *Thermosynechoccus elongatus* (*T. elongatus*) was recently found to contain 28 closely related group II introns (Nakamura et al., 2002). *T.*

elongatus is a cyanobacterium with an optimal growth temperature of 55°C . The closest known relative of the *T. elongatus* introns is the previously characterized *Escherichia coli* intron EcI5, with about 50% sequence identity (Mohr et al., 2010; Zhuang et al., 2009). The EcI5 and *T. elongatus* introns belong to the chloroplast-like IIB1 subclass (Dai and Zimmerly, 2002b; Mohr et al., 2010). Seventeen of the *T. elongatus* introns do not encode proteins (TeI3a– t), while eight of the introns contain open reading frames (ORF) that encode IEPs (TeI4a- h) (Mohr et al., 2010). Five of the ORFs are interrupted by the insertion of an ORF-less intron, also known as a twintron (Dai and Zimmerly, 2002a; Mohr et al., 2010).

Most bacterial genomes contain only one or two group II introns (Dai and Zimmerly, 2002a). The group II introns from *T. elongatus* proliferated to high numbers using several mechanisms (Mohr et al., 2010). First, the introns' EBS sequence diverged into six different families allowing the introns to target different sites. Second, the IEPs likely evolved to have less intron specificity, enabling the IEP to mobilize ORF-less introns. Lastly, higher temperatures promoted DNA strand separation, allowing the target site to be recognized almost entirely by intron base pairing (Mohr et al., 2010).

1.4 Reverse transcriptase activity of TeI4c

Group II introns from *T. elongatus* encode heat-stable proteins, whose reverse trancriptase activity could possibly be utilized commercially for applications involving cDNA synthesis (Vellore et al., 2004). cDNA synthesis can be hindered by RNA secondary and tertiary structures, such as helices and kinks that form in the RNA; however, increasing the reaction temperature can destabilize these structures. Previously,

it was shown that TeI4c is thermally stable and active at high temperatures when fused to maltose-binding protein (MBP) (S. Mohr, unpublished data). However, it remained uncertain whether other purification tags could be utilized for biochemical applications. This work examines the effect that N utilization substance A (NusA) (Nallamsetty and Waugh, 2006), a solubility-enhancing protein, has on TeI4c reverse transcriptase activity.

Of the eight IEPs, TeI4c has the highest reverse transcriptase activity (S. Mohr, unpublished data). TeI4c was found to be stable and active when N-terminally fused to MBP through a short five-alanine amino acid linkage or rigid fusion (Smyth et al., 2003). MBP is a large affinity tag that increases the solubility and stability of proteins (Kapust and Waugh, 1999; Nallamsetty and Waugh, 2006). The effect of the endonuclease and DNA binding domains on the reverse transcriptase activity of TeI4c was unknown and is the subject of this study.

1.5 DNA target site recognition for TeI4c

TeI4c recognizes 1-2 nucleotides at the 5'-exon of the DNA target site (Mohr et al., 2010). This interaction is required for reverse splicing into double stranded DNA; however, the IEP does not recognize any nucleotide in the 3'- exon (Mohr et al., 2010). For other characterized group II introns, the IEP is required to recognize 3'-exon nucleotides for endonuclease cleavage (Mohr et al., 2000). The lack of recognition of 3' exon nucleotides for retrohoming by TeI4c could be explained in several ways: (1) the 5'-EBS/IBS interactions may be sufficient for site-specific bottom-strand cleavage, (2) endonuclease cleavage may not be site-specific, or (3) the introns may not rely on endonuclease cleavage and use endonuclease-independent mobility (Mohr et al., 2010).

This work examines these possibilities by studying the mobility of TeI4c with and without the endonuclease domain (ΔEn) to determine whether TeI4c utilizes endonuclease-independent mobility predominantly. Additionally, the endonuclease activity was examined in an *in vitro* biochemical assay to investigate whether TeI4c has site-specific or non-site-specific endonuclease activity.

1.6 Protein structure of HIV-RT and group II intron-encoded proteins

To date, the structure of a group II intron-encoded protein has not been determined. However, the reverse transcriptase domain of group II intron-encoded proteins is homologous to that of retroviral reverse transcriptases, such as HIV-RT, whose structure has been determined (Kohlstaedt et al., 1992; Michel and Lang, 1985; Zimmerly et al., 2001). HIV-RT is a heterodimer consisting of the subunits, p66 and p51 (Figure 1.7). Each subunit can be described as having fingers, palm, thumb, and connector domains. The fingers and thumb of the reverse transcriptase form a cleft with the palm and RT active site at the base (Kohlstaedt et al., 1992; Sarafianos et al., 2009). Utilizing the previously determined structure of HIV-RT (Kohlstaedt et al., 1992), a three-dimensional model of LtrA was proposed by threading the aligned sequence of LtrA's RT, X, and DNA binding domains onto the structure of HIV-1 RT (Blocker et al., 2005). It is predicted that group II intron IEPs form an active site similar to HIV-RT (Blocker et al., 2005).

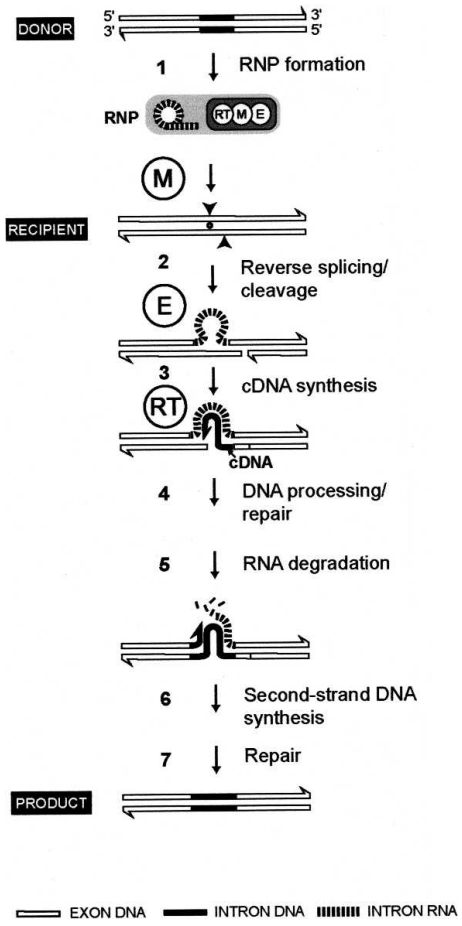
One aim of this work was to crystallize TeI4c ΔEn. Two strategies utilized to increase the chances of crystal formation are crystallization using a carrier protein and surface entropy reduction (SER) (Derewenda, 2004). Using a large carrier protein, such

as maltose binding protein, can increase expression levels, improve protein solubility, aid in protein folding, and encourage the formation of a crystal lattice (Kapust and Waugh, 1999; Smyth et al., 2003). SER addresses the problem that crystallization can be inhibited by the entropic behavior of large hydrophilic side chains on the protein surface. The SER technique changes patches of surface-exposed large, charged residues to small nonpolar amino acids. This reduces surface entropy and may enable crystal contacts (Avbelj and Fele, 1998). Surface entropy mutations of maltose binding protein have been designed previously and were used in this study (Moon et al., 2010) (Table 1.1). The two strategies, carrier proteins and surface entropy reduction, have been explored in this study as a means to induce crystallization of TeI4c.

1.7 Overview of thesis research

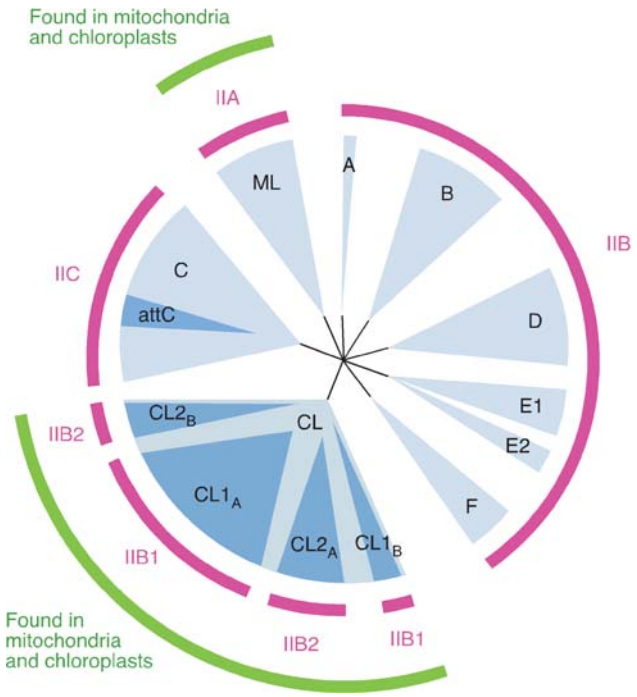
This thesis work examines the group II intron-encoded proteins from *Thermosynechoccus elongatus*. I addressed the following topics: 1) biotechnology applications of group II intron-encoded reverse transcriptases, 2) characterization of the intron-encoded proteins' DNA endonuclease domain and endonuclease-independent retrohoming, and 3) the crystallization of group II intron-encoded proteins. The results of this work are summarized in the following chapters.

Figure 1.1: Retrohoming pathway for the L1.LtrB intron



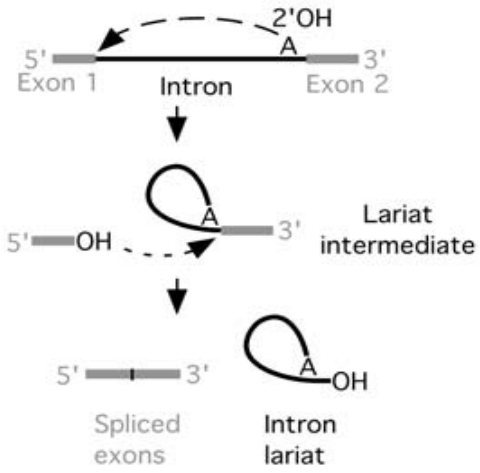
Group II intron retrohoming occurs through the following steps: (1) Transcription of the intron from the donor DNA, and protein-assisted splicing to form ribonucleoprotein particles (RNPs). (2) Cleavage of the top strand through reverse splicing of intron into the target DNA, followed by bottom-strand cleavage by the endonuclease (En) domain. (3) cDNA synthesis of the intron using the nicked second-strand as a primer. (4-7) DNA processing, RNA degradation, second-strand DNA synthesis, and DNA repair. Adapted from Smith et al. (2005).

Figure 1.2: Phylogeny of group II intron-encoded proteins and RNA structural classes



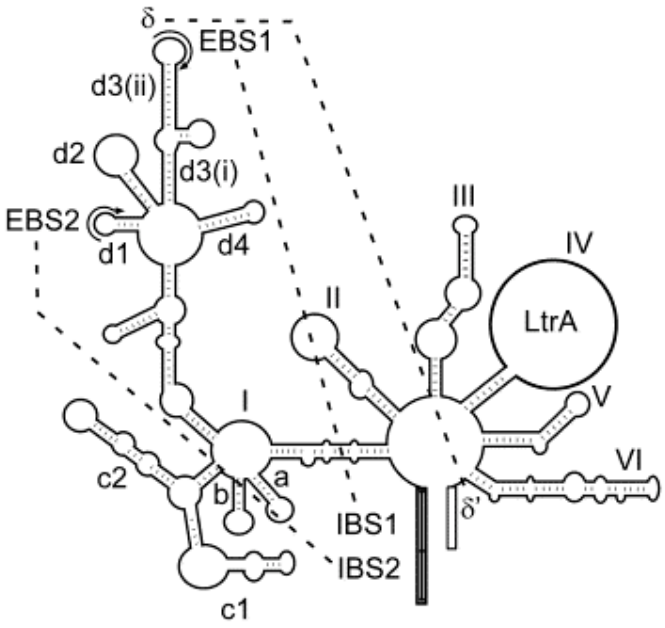
Group II intron-encoded proteins are divided into nine classes (depicted in blue) termed mitochondrial-like (ML), chloroplast-like (CL) 1 and 2, bacterial A-F. Each intron-encoded protein is associated with an intron RNA structural class: IIA1, IIB1, IIB2, IIC, IIB-like, and IIA/B (depicted in pink). Adapted from Lambowitz and Zimmerly (2010).

Figure 1.3: Group II intron splicing mechanism



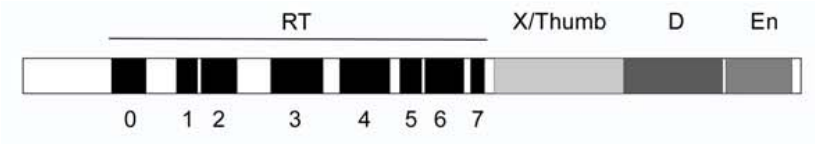
Group II introns splice though two transesterification reactions. First, a 2' OH from the internal adenosine of domain DVI acts as a nucleophile, attacking the 5'-splice site. This results in a lariat/3'-exon intermediate. Second, the 3' OH from the cleaved 5'-exon acts a nucleophile, attacking the 3'-splice site. This results in exon ligation and the formation of the intron lariat RNA. Adapted from Lambowitz and Zimmerly (2004).

Figure 1.4: Group II intron RNA secondary structure



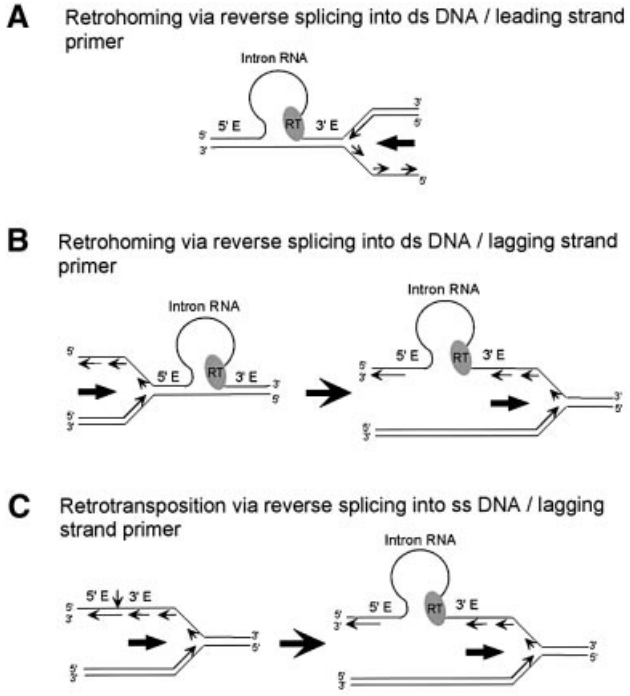
The secondary structure of L1.LtrB consists of six double-helical domains (DI- DVI) radiating from a central wheel. Domain IV contains the open reading frame encoding LtrA. Domain I contains the exon-binding sites (EBS1, EBS2) which base pair with the 5'-exon intron-binding site sequences (IBS1, IBS2). δ base pairs with the first 1-3 nucleotides of the 3'-exon, δ'. Adapted from Perutka et al. (2004).

Figure 1.5: Schematic of a group II intron-encoded protein



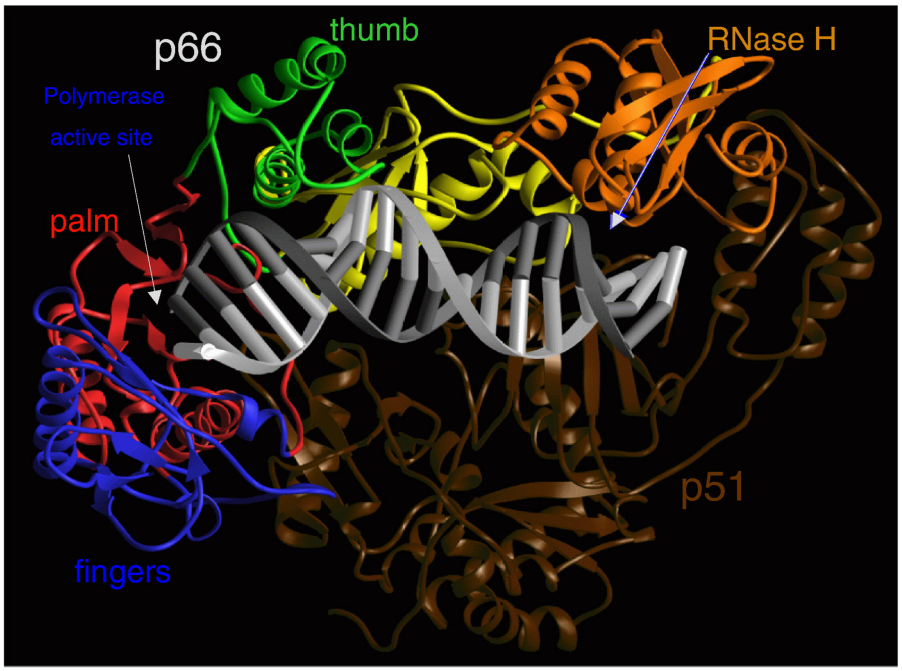
Group II intron-encoded proteins are multidomain proteins consisting of the following domains: reverse transcriptase (RT), maturase (X), DNA binding (D), and endonuclease domains (En). The reverse transcriptase domain consists of the conserved sequence blocks RT1- RT7, which are characteristic of the finger and palm regions of retroviral RTs. The RT0 sequence block is characteristic of RTs encoded by non-LTR retrotransposons. The maturase domain (X) corresponds to the thumb in retroviral RTs. The DNA endonuclease domain (En) is used in bottom-strand cleavage. Adapted from Mohr et al. (2010).

Figure 1.6: Endonuclease-independent retrohoming mechanisms



Models for group II intron mobility using leading and lagging-strands at the replication fork to prime reverse transcription. Retrohoming is shown through reverse splicing into double-stranded DNA and using the (A) leading strand or (B) lagging strand at the replication fork as a primer. (A) The leading strand is directly used as a primer before the replication fork passes. (B) The lagging strand is used a primer after the replication fork has traversed the inserted intron RNP. (C) The group II intron reverse splices into single-stranded DNA using a lagging strand as a primer. Adapted from Zhong and Lambowitz (2003).

Figure 1.7: Structure of HIV-RT in complex with nucleic acid



Ribbon diagram of HIV-RT in complex with nucleic acid. HIV-RT is a heterodimer consisting of the subunits p66 and p51. The p66 subunit: fingers, palm, thumb, connection and RNase H domains are depicted in blue, red, green, yellow, and orange, respectively. The p51 subunit is depicted in brown. The template nucleic acid is depicted in gray. The fingers and thumb of the reverse transcriptase form a cleft with the palm and active site at the base. Adapted from Sarafianos et al. (2009).

Table 1.1: Surface entropy mutations for maltose binding protein

Table II. *SER Mutations Present in the MBP-SER Cassettes*

Vector	SER mutations
pMALX(A)	D82A/K83A
pMALX (B)	E172A/N173A
pMALX(C)	D82A/K83A/K239A
pMALX(D)	E172A/N173A/K239A
pMALX(E)	D82A/K83A/E172A/N173A/K239A

TeI4c was N-terminally fused to maltose binding protein (MBP). Five vectors (A-E) were developed with different surface entropy reduction mutations. Adapted from Moon et al. (2010).

Chapter 2: Group II intron-encoded reverse transcriptases for biotechnology applications

Reverse transcriptases (RTs) are used in biotechnology to copy RNA into DNA. However, the RNA template can form secondary or tertiary structures, such as kinks or helices, that hinder cDNA synthesis. Increasing the reaction temperature destabilizes secondary and tertiary structures and minimizes mispriming, thereby optimizing the reverse transcriptase reaction. Currently, most commercially available reverse transcriptases are derived from avian myeloblastosis virus (AMV) and marine leukemia virus (M-MLV). The commonly used thermostable RT, SuperScript III is a derivative of M-MLV RT and is active at temperatures up to 55°C (Potter et al. 2003).

Group II intron-encoded proteins have inherently high processivity and fidelity, making them ideal candidates for a commercial reverse transcriptase. *In vivo*, the IEP reverse transcribes the highly structured intron RNA, which is ~ 2-2.5 kb in length. Additionally, *in vitro* the TeI4c RT IEP has also been shown to be processive (S. Mohr, unpublished data). The L1.LtrB RT has an *in vivo* error frequency of 2.2×10^{-5}, which is lower than the error frequencies of commercially available AMV RT and M-MLV RT (range of 4.8×10^{-5}) (Bakhanashvili and Hizi, 1993; Conlan et al., 2005).

The TeI4c protein was found to be active and thermostable when maltose binding protein (MBP) was N-terminally fused to TeI4c through a short rigid fusion (RF) of five alanines. The MBP-RF- TeI4c protein had optimal reverse transcriptase activity at 61°C (S. Mohr, unpublished results). This work examines the possibility that other solubility tags, such as NusA (N utilization substance protein) could be used as an alternative to

maltose binding protein as a commercially available RT fusion protein. NusA was N-terminally fused to TeI4c through a rigid fusion and the reverse transcriptase activity was determined. This work compares the RT activity of TeI4c when fused to NusA or MBP.

2.1 Reverse transcriptase activity of TeI4c when fused to NusA

A vector was constructed with TeI4c N-terminally fused to NusA through a rigid fusion (RF) of five alanine residues and a C-terminal His6 tag used for purification (Figure 2.1). The protein was expressed in *E. coli* and purified by a procedure that involves polyethylenimine (PEI) precipitation of nucleic acids followed by nickel affinity and heparin-Sepharose chromatography. The protein was dialyzed into buffer with 50% glycerol and flash frozen. Through polyacrylamide gel electrophoresis the protein was determined to be > 95% pure, with a yield of ~ 0.3 mg/ L.

The protein was assayed for reverse transcriptase activity from 25° to 77°C using an artificial substrate, which consisted of a poly(rA) template annealed to a 42-nucleotide oligo(dT) primer. The activities of TeI4c with either the NusA or MBP tag were assayed by quantifying the polymerization of ^{32}P- dTTP. NusA-RF-TeI4c was active at high temperatures with an optimum temperature range of 41° to 65° C (Figure 2.2). The reverse transcriptase activity significantly decreased at 73°C. Compared to MBP-RF-TeI4c, NusA-RF-TeI4c was ~ 3 fold less active, though both constructs have a similar optimal RT temperature range from 41°C to 65°C.

2.2 Discussion

When expressed as either a NusA or MBP N-terminal fusion the TeI4c protein was thermostable and active. The NusA-TeI4c protein fusion was somewhat less active

than the protein fused to MBP. Thus, different solubility tags can vary in their effectiveness at stabilizing the reverse transcriptase of group II introns IEPs. At present, the maltose binding protein tag provides the best, known solubility tag for TeI4c biotechnology applications.

2.3 Materials and Methods

Recombinant plasmids

The TeI4c RT protein was fused to an N-terminal NusA tag and MalE tag via a rigid five-alanine linker. The pMalE-RF-TeI4c was constructed previously (S. Mohr, unpublished data). pNusA-RF-TeI4c-His6 was constructed by PCR amplification of pMalE-RF-TeI4c with primers that append the restriction enzymes SacII and KpnI. The PCR product was then cloned into the corresponding sites of pET-50b(+) (Novagen), a vector for expressing proteins with NusA fusions. Through PCR mutagenesis the last two charged residues (D and E) of NusA were replaced with alanines, and the existing linker (NICWFGDEATSGSGH6) was replaced by the rigid fusion linker (NICWFGAAAAA). Two N-terminal His6 tags were removed, and a C-terminal His6 tag was added to facilitate purification.

Expression and purification

pNusA-RF-TeI4c was expressed in *E. coli* ScarabXpress T7lac competent cells (Scarab Genomics). Cells were grown in Terrific Broth (TB) until an optical density of ~ 1.4 – 1.6 (OD_{600}), were induced with 0.5 mM IPTG, and grown for an additional 48 h at 18° C. Cells were pelleted and resuspended in nickel buffer A (20 mM Tris-HCl pH 7.5, 500 mM KCl, 30 mM imidazole, 10% glycerol). The cells were then disrupted through

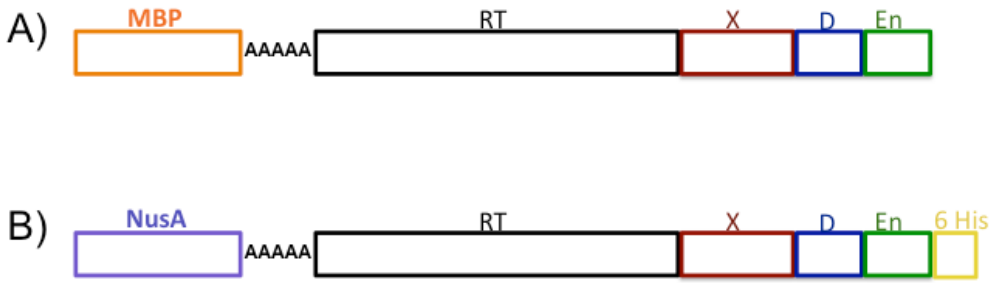
three cycles of freeze/ thaw with 1 mg/ml lysozyme, and sonicated (Branson 450 Sonifer, Branson Ultrasonics, Danbury CT; two 45 sec bursts at amplitude of 60%). Nucleic acids were precipitated from the lysate by adding a final concentration of 0.2% polyethyleneimine followed by centrifugation at 12,000 x g for 15 min. The supernatant was then purified using a 5-ml nickel-Sepharose (GE Healthcare) column that had been equilibrated with nickel buffer A. The protein was eluted with nickel buffer A containing 500 mM imidazole. The protein sample was then applied to two connected 1-ml heparin-Sepharose (GE Healthcare) columns that had been equilibrated in heparin-Sepharose buffer (20 mM Tris-HCl, pH 7.5, 100 mM KCl, 1 mM DTT, 1 mM EDTA, and 20% glycerol) and eluted with a 20-column volume gradient of 0.1 to 1.5 M KCl. The protein eluted around 800 mM KCl, and the protein was pooled and dialyzed overnight into storage buffer (20 mM Tris-HCl pH 7.5, 0.5 M KCl, 1 mM EDTA, 1 mM DTT, 50% glycerol).

Reverse transcriptase assay

The reverese transcriptase activity of the NusA-RF-TeI4c and MBP-RF-TeI4c proteins were measured from 25° to 77°C using an artificial substrate, consisting of a poly(rA) template annealed to a 42-nucleotide oligo(dT) primer. The reverse transcriptase activity was assayed by quantifying the polymerization of ^{32}P-dTTP. The RT (NusA-RF-TeI4c, 100 nM, MBP-RF-TeI4c, 50 nM) in RT buffer (75 mM KCl, 10 mM $MgCl_2$, 20 mM Tris-HCl pH 7.5, 1 mM DTT) was pre-incubated with 100 nM of the poly(rA)/oligo$(dT)_{42}$ substrate at the reaction temperature. The reaction was then initiated by adding 5 μCi of [α-^{32}P]-dTTP (3,000 Ci/mmol;Perkin Elmer, Waltham MA). The

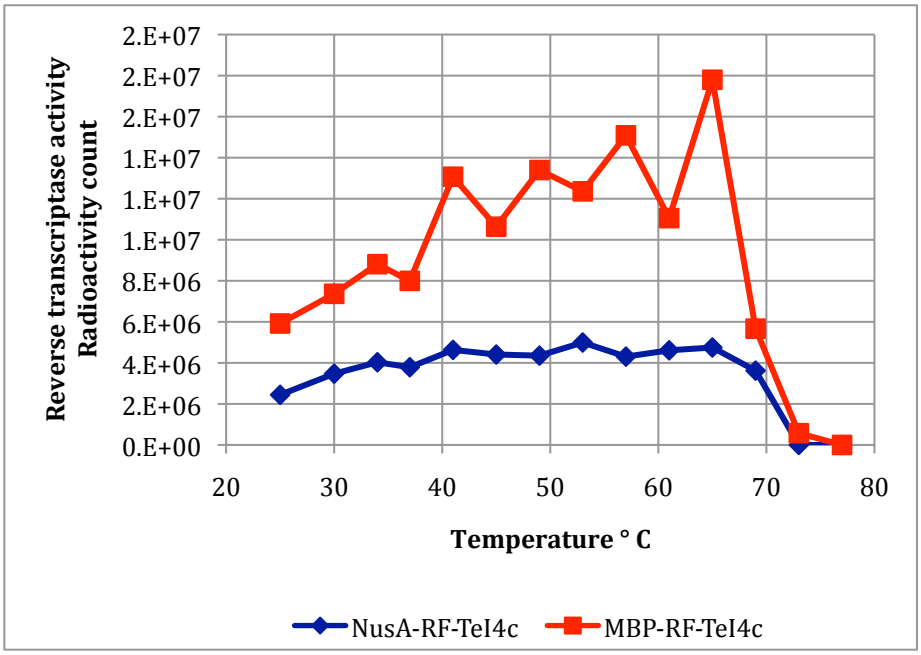
reactions were incubated for times within the linear range (NusA-RF-TeI4c 2 min, MBP-RF-TeI4c 1.5 min) and stopped by adding a final concentration of 250 mM EDTA. The reaction products were then spotted onto Whatman DE81 chromatography paper (GE Healthcare), washed 3 times in 0.3 M NaCl and 0.03 M sodium citrate, and scanned with a PhosphorImager (Typhoon Trio Variable Mode Imager; GE Healthcare). Bound radioactivity was quantified using ImageQuant (GE Healthcare).

Figure 2.1: Schematics of MBP-RF-TeI4c and NusA-RF-TeI4c



TeI4c was N-terminally fused to (A) maltose binding protein or (B) N utilization substance A through a five alanine rigid fusion (RF). (B) The C-terminus of NusA-RF-TeI4c has a 6His tag used for purification.

Figure 2.2: Reverse transcriptase activity of NusA-RF-TeI4c and MBP-RF-TeI4c



Reverse transcriptase assay of the NusA-RF-TeI4c and MBP-RF-TeI4c proteins from 25° to 77°C using poly(rA)/oligo(dT)$_{42}$ as a substrate. The activity plotted was corrected for background bound radioactivity.

Chapter 3: Characterization of the DNA binding and DNA endonuclease domains of TeI4c RT

Deletion of the endonuclease domain of the L1.LtrB group II intron RT, the LtrA protein, abolishes not only bottom-strand cleavage but also reverse transcriptase activity, even using simple artificial template-primer substrates (San Filippo and Lambowitz, 2002). However, not all IEPs require the endonuclease domain for reverse transcriptase activity. EcI5, the closest known relative to *T. elongatus*, is one such example (Zhuang et al., 2009). This work examines the effect of deleting the C-terminal DNA binding and endonuclease domains on the reverse transcriptase activity for the group II intron-encoded protein TeI4c.

To examine the effect of deleting the DNA binding and endonuclease domains of TeI4c, three C-terminal truncations were constructed denoted to identify the first truncated amino acid residues: Δ439, Δ484, and Δ502. The truncations were made in the pMalE-RF-TeI4c vector background, which encodes a N-terminal maltose binding protein fused through a short rigid fusion (RF) to the TeI4c RT. The reverse transcriptase activity of the truncations was determined through two RT assays, a poly(rA)/oligo(dT) assay and a primer extension assay.

3.1 Sequence alignment of TeI4c with LtrA

The secondary structure and domains of LtrA have been previously predicted (Blocker et al., 2005; San Filippo and Lambowitz, 2002). Using this information, the secondary structure and domain boundaries of the TeI4c RT were predicted by aligning TeI4c with LtrA (Figure 3.1). The sequence alignment of LtrA and TeI4c was made using

the multiple alignment program ClustalW2 (Chenna et al., 2003). The group II intron-encoded proteins contain eight conserved reverse transcriptase sequence motifs, RT0 – RT7 (Zimmerly et al., 2001). The sequence motifs RT1 – RT7 are conserved in retroviral RTs, while the RT0 sequence is characteristic of non-LTR-retrotransposon RTs (Malik et al., 1999; Zimmerly et al., 2001). The alignment of LtrA and TeI4c shows conserved amino acids within the RT0 –RT1 sequence elements. In LtrA, seven α-helices are predicted downstream of the reverse transcriptase domain. Figure 3.1 includes these predicted α-helices in the alignment.

Through examination of the sequence alignment, three C-terminal truncations were designed: Δ439, Δ484, and Δ502. All three truncations were constructed with TeI4c N-terminally fused to maltose binding protein through a rigid fusion (Figure 3.2). All the truncations removed the C-terminal endonuclease domain. The TeI4c Δ502 RT removed only the endonuclease domain. The TeI4c Δ484 RT removed the α-helix, αD, from the DNA binding domain, and the TeI4c Δ439 RT removed the entire DNA binding domain.

3.2 Reverse transcriptase activity of C-terminal truncations

The C-terminal truncations were constructed through PCR QuikChange site-directed mutagenesis of the pMALE-RF-TeI4c expression plasmid using oligonucleotides that contained the desired deletion (Wang and Malcolm, 1999). The full-length and truncated TeI4c constructs were expressed in *E. coli* and purified through chromatography.

The full-length and truncated TeI4c proteins were assayed for reverse transcriptase activity *in vitro* at 60°C using an artificial poly(rA)/oligo$(dT)_{42}$ substrate. The C-terminal truncations, Δ502 and Δ484, displayed reverse transcriptase activity (Figure 3.3). The TeI4c Δ439 RT, which removed the entire DNA binding and endonuclease domains, displayed little or no reverse transcriptase activity. Whereas, the TeI4c Δ484 RT displayed ~30% of full-length activity, the TeI4c Δ502 RT was surprisingly ~ 3 times more active than full-length TeI4c protein at 60° C.

Encouraged by these results, I then assayed the reverse transcriptase activity of TeI4c Δ502 at temperatures ranging from 25° to 77°C (Figure 3.4). The TeI4c Δ502 RT had a temperature profile similar to the full-length TeI4c protein with the TeI4c Δ502 protein displaying ~ 2 – 3 fold more activity than the full-length protein. The TeI4c Δ502 RT had the greatest activity at temperatures from 61° to 65°C, with its peak activity at 61°C.

The processivity of the truncations were examined *in vitro* using a 531-nt transcript. The transcript was synthesized from an AflIII-digested plasmid, pBS KS(+), that was transcribed and then annealed to a ^{32}P-labeled 37-nt primer (AflIII). The labeled transcript and truncated proteins were incubated at 60 °C for 5 min and the products were analyzed by electrophoresis in a denaturing 6% polyacrylamide gel (Figure 3.5). The TeI4c Δ502 and Δ484 RTs were processive enzymes, synthesizing full-length cDNAs. Similar to the poly(rA)/oligo(dT) assay, the TeI4c Δ439 protein had little to no reverse transcriptase activity. The TeI4c Δ484 RT had ~ 23% of the full-length protein activity, and the TeI4c Δ502 protein had ~ 1.5 fold more activity than the full-length protein.

3.3 Discussion

The endonuclease domain of TeI4c is not required for reverse transcriptase activity. Surprisingly, the TeI4c protein has significantly more RT activity *in vitro* without the endonuclease domain (Δ502). Unlike LtrA, TeI4c does not require the endonuclease domain to interact with the RT domain for activity (San Filippo and Lambowitz, 2002). Deletion of the entire DNA binding domain abolishes reverse transcriptase activity, and partial deletion of the DNA binding domain significantly decreases RT activity. The DNA binding and endonuclease domains had no affect on the RT's processivity.

3.4 Materials and Methods

Recombinant plasmids

The expression plasmids pMalE-RF-TeI4c Δ502, pMalE-RF-TeI4c Δ484, and pMalE-TeI4c Δ439 were constructed through truncation of the pMalE-RF-TeI4c vector. The vector was truncated through a variation of QuikChange site-directed mutagenesis with large oligonucleotides that flanked the desired deletion, but did not include the truncation (Wang and Malcolm, 1999).

Expression and purification

The plasmids pMalE-RF-TeI4c Δ502, pMalE-RF-TeI4c Δ484, and pMalE-TeI4c Δ439 were expressed in *E. coli* Rossetta 2 competent cells (Novagen, EMD Biosciences, Gibbstown NJ). The cells were grown in Terrific Broth (TB) at 37°C until mid-log phase (OD_{600} ~1.4- 1.6), induced with 500 μM IPTG, and grown for an additional 48h at 18°C. The cells were then pelleted and resuspended in amylose A buffer (20 mM Tris-HCl pH

7.5, 500 mM KCl, 1 mM DTT, 1 mM EDTA, 20% glycerol). The cells were disrupted through three freeze/thaw cycles with lysozyme (1 mg/ml of lysate), and sonicated (Branson 450 Sonifer, Branson Ultrasonics, Danbury CT; two 45 sec bursts at amplitude of 60%). The cells were centrifuged at 14,000 x g for 40 minutes. Nucleic acids were precipitated from the lysate by adding a final concentration of 0.2% polyethyleneimine, followed by centrifugation at 12,000 x g for 15 min. The supernatant was then purified using a 10-ml amylose column (Amylose High Flow, New England Biolabs) that had been equilibrated in amylose A buffer. The protein was washed with buffer containing 1.5 M KCl, and then eluted with amylose buffer containing 10 mM maltose. The protein peak was pooled and applied to a heparin- Sepharose column (GE Healthcare) that had been equilibrated in heparin-Sepharose buffer (20 mM Tris-HCl pH 7.5, 100 mM KCl, 1 mM DTT, 1 mM EDTA, and 20% glycerol) and the protein was eluted with a 20-column volume gradient of 0.1 to 1.5 M KCl. The proteins eluted around 800 mM KCl, and were pooled and dialyzed overnight into storage buffer (20 mM Tris-HCl pH 7.5, 0.5 M KCl, 1 mM EDTA, 1 mM DTT, 50% glycerol).

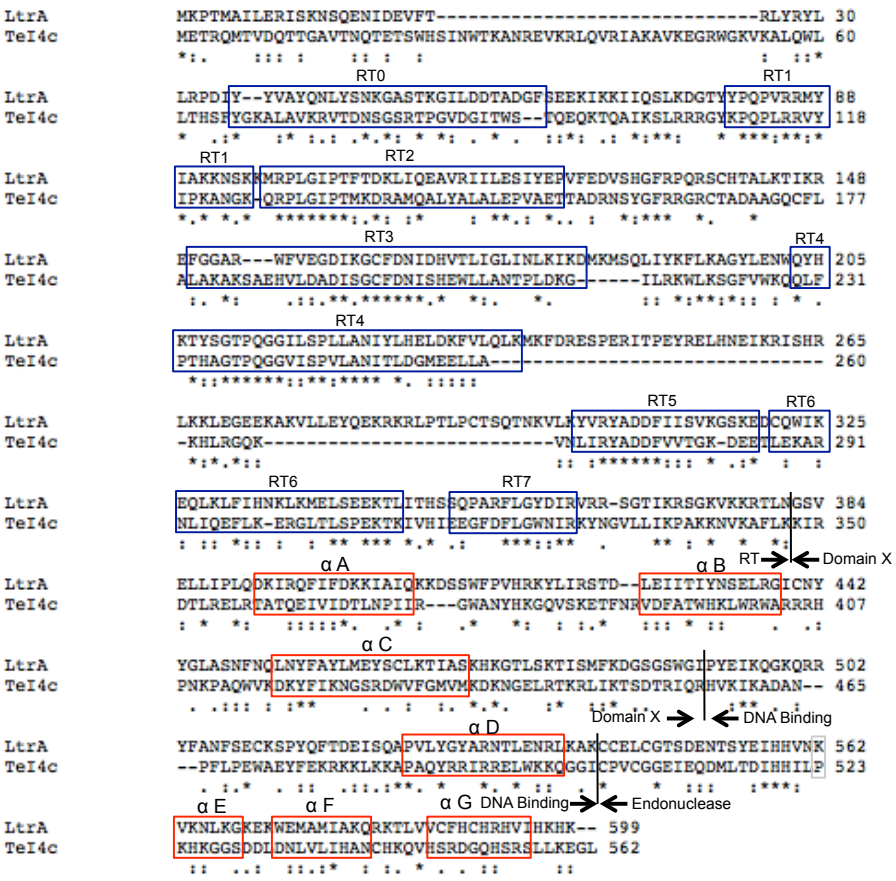
Reverse transcriptase assay

The activity of the MBP-RF-TeI4c, MBP-RF-TeI4c Δ502, MBP-RF-TeI4c Δ484, and MBP-RF-TeI4c Δ439 proteins (100 nM) was measured at 60°C with the poly(rA)/oligo$(dT)_{42}$ substrate (100 nM). Additionally, the activity of the MBP-RF-TeI4c and MBP-RF-TeI4c Δ502 proteins (50 nM) was measured from 25°- 77°C. The RTs were pre-incubated with the poly(rA)/oligo$(dT)_{42}$ substrate at the reaction temperature, and the reaction was initiated by adding 5 μCi of [α-^{32}P]-dTTP (3,000 Ci/mmol; Perkin

Elmer, Waltham MA). The reactions were incubated for 1.5 min, and terminated by adding EDTA to a final concentration of 250 mM. The products were spotted onto Whatman DE81 chromatography paper (GE Healthcare), washed three times in 0.3 M NaCl and 0.03 M sodium citrate, and scanned with a PhosphorImager (Typhoon Trio Variable Mode Imager; GE Healthcare).

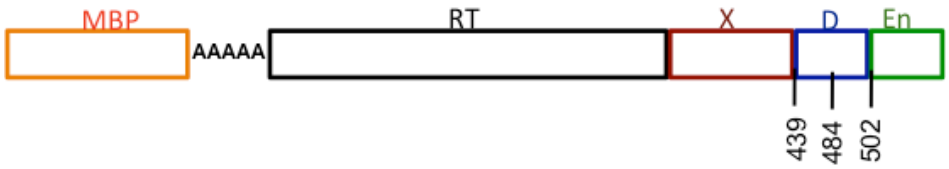
Reverse transcriptase activity was also determined through a primer extension reaction using a pBSAflIII 531-nt RNA template that was annealed to the oligonucleotide AflIII (AflIII primer: 5'- CCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCG). The RNA template was made by digesting the plasmid pBluescript (+) KS with AflIII and then transcribed using T7 Megashortscript (Ambion, Austin, TX). The oligonucleotide AflIII was designed to be complementary to the 3'- end of the RNA template. The primer was ^{32}P-labeled with T4 polynucleotide kinase (New England Biolabs) and annealed to the RNA template at 82°C for 2 min. The RT (100 nM) and RNA substrate (30 nM) in primer extension buffer (750 mM KCl, 100 mM $MgCl_2$, 200 mM Tris-HCl pH 7.5, 1 mM DTT) were pre-incubated for one min at 60°C. The reaction was initiated by adding pre-warmed to reaction temperature, dNTPs to a final concentration of 2 mM. The reaction was incubated for 5 min, and terminated by adding a final concentration of 250 mM EDTA and 0.1% SDS. The product was phenol-extracted and loaded onto a 6% polyacrylamide denaturing gel.

Figure 3.1: Sequence alignment of LtrA and TeI4c



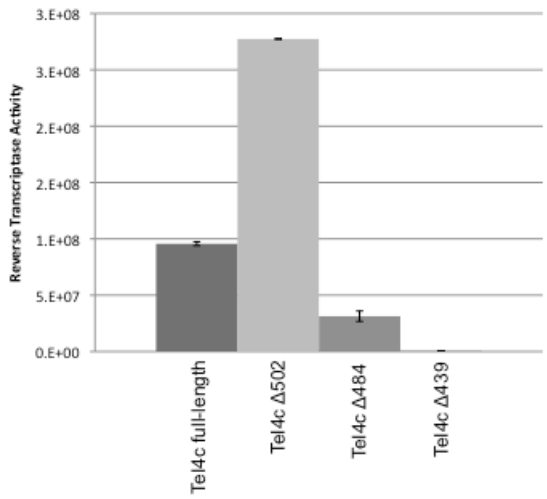
Sequence alignment of LtrA and TeI4c using ClustalW2 (Chenna et al., 2003). The conserved reverse transcriptase domains, RT0 –RT7, are indicated in blue boxes. Alpha helices, αA- αG, are indicated in red boxes.

Figure 3.2: Schematic of TeI4c C-terminal truncations



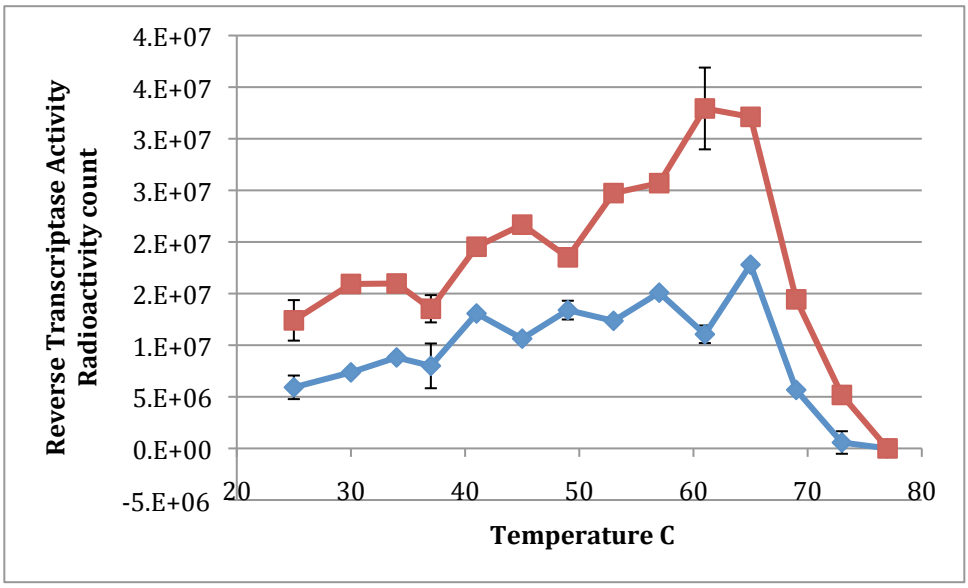
TeI4c was N-terminally fused to maltose binding protein (MBP) through a rigid linker of five alanines. Three C-terminal truncations were constructed (a) Δ439, (b) Δ484, and (c) Δ502. All truncations removed the entire endonuclease domain.

Figure 3.3: Poly(rA)/oligo(dT) reverse transcriptase activity of TeI4c truncations



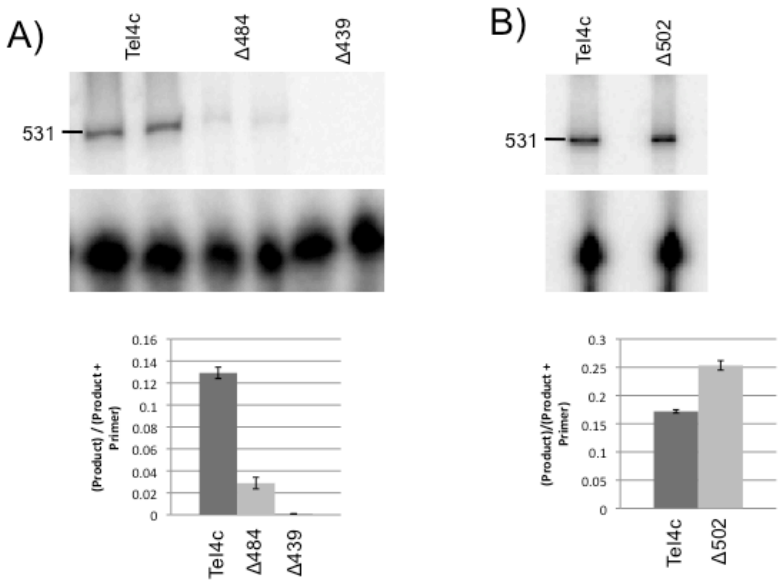
The reverse transcriptase activity of full-length and truncated TeI4c protein (Δ502, Δ484, and Δ439; 100 nM) were assayed using an artificial poly(rA)/oligo(dT) substrate at 60°C for 1.5 min, which is within linear range.

Figure 3.4: Poly(rA)/oligo(dT) reverse transcriptase activity of TeI4c Δ502 from 25-77°C



The reverse transcriptase activity of the MBP-RF-TeI4c and MBP-RF-TeI4c Δ502 proteins (50 nM) were measured using a poly(rA)/oligo(dT)$_{42}$ substrate (100 nM) from 25°C to 77°C for 1.5 min. The activity plotted was corrected for background bound radioactivity.

Figure 3.5: Primer extension reverse transcriptase activity of full-length and truncated TeI4c protein



The top and bottom panels show the full-length product and unextended primer, respectively. The bar graphs below show the percentage of primer extended to full-length cDNA. (A) Reaction shown in duplicate: TeI4c, TeI4c Δ484, and TeI4c Δ439. (B) Reactions: TeI4c and TeI4c Δ502.

Chapter 4: Biochemical characterization of the endonuclease domain of TeI4c and its utilization in retrohoming

LtrA and other characterized group II intron-encoded proteins require that the IEP recognize the 5'-exon and 3'-exon for reverse splicing and bottom-strand cleavage, respectively (Guo et al., 1997; Saldanha et al., 1999). A previous study examined TeI4c's ability to recognize the DNA target site. The TeI4c IEP was found to recognize specific nucleotide residues in the 5'-exon region of the DNA target site (Mohr et al., 2010; Mohr et al., 2000). However, the TeI4c IEP did not recognize a nucleotide in the 3'- exon as required by other characterized group II introns for bottom-strand DNA cleavage (Mohr et al., 2010). The lack of 3'-exon recognition suggests several possible mechanisms that could be used in TeI4c mobility: (1) 5'-EBS/IBS interactions may be sufficient for endonuclease cleavage, (2) bottom-strand cleavage is not site-specific, or (3) the intron may not rely on endonuclease cleavage and use endonuclease-independent mobility for retrohoming (Mohr et al., 2010). This chapter examines the ability of the TeI4c IEP to use endonuclease-independent mobility through a plasmid-based mobility assay, which compares the retrohoming efficiency of full-length TeI4c protein to TeI4c Δ502 protein. Additionally, this chapter examines the biochemical activity of the TeI4c protein through an *in vitro* endonuclease assay using reconstituted ribonucleoproteins (RNPs) to cleave ^{32}P end-labeled DNA target-site substrates.

The closest known relative of *T. elongatus* is EcI5 (Mohr et al., 2010). Although EcI5 contains an endonuclease domain, it can also use En-independent retrohoming mechanisms for retrohoming (Zhuang et al., 2009). As shown in the previous chapter,

TeI4c does not require the endonuclease domain for reverse transcriptase activity. These findings suggest TeI4c may be similar to EcI5 and not require the endonuclease domain for retrohoming. It is feasible that TeI4c utilizes En-independent retrohoming, entirely or partially, for mobility.

The endonuclease activity of several group II intron-encoded proteins has been previously characterized (Lambowitz and Zimmerly, 2010). The top strand is cleaved as the intron reverse splices into the target site. All characterized group II introns cleave the top strand at the intron insertion site (Lambowitz and Zimmerly, 2010; Matsuura et al., 1997; Zhuang et al., 2009). However, cleavage of the bottom strand by the endonuclease domain varies with the group II intron. LtrA cleavage occurs at position +9 of the 3' exon (Matsuura et al., 1997). The endonuclease cleavage for the yeast mitochondrial DNA group II introns, aI1 and aI2, occurs at position +10 of the 3' exon (Lambowitz and Zimmerly, 2010). This work characterizes the endonuclease activity of the TeI4c protein.

4.1 Endonuclease-independent retrohoming of TeI4c

Mobility assays were performed using an *in vivo* plasmid-based mobility assay (Guo et al., 2000; Karberg et al., 2001). The mobility assay uses two vector constructs to measure retrohoming events: a donor vector containing an ORF-less intron RNA upstream of the intron-encoded protein, and a recipient vector containing the DNA target site (Figure 4.1). The chloramphenicol-resistant donor plasmid uses a T7Lac promoter to express the ORF-less intron and the IEP. The intron RNA carries a T7 promoter sequence in DIV. The recipient plasmid contains the target sites (ligated E1-E2 sequences) upstream of a promoterless tetracycline-resistance (tet^R) gene. The assay is performed in

E. coli HMS174(DE3) cells, which contain an IPTG-inducible T7 RNA polymerase that induces intron expression with IPTG. As the intron carrying the T7 promoter integrates into the target site, the tetracycline-gene is activated, allowing for the selection of retrohoming events (Guo et al., 2000; Mohr et al., 2010).

This work examines the possibility that TeI4c retrohomes through endonuclease-independent pathways. En-independent pathways use nascent strands at DNA replication forks as primers for reverse transcription. The primer can be from either a leading (LEAD) or lagging (LAG) DNA strand. The intron encoded protein, LtrA has previously shown to have a bias for using the nascent leading strand primer during retrohoming (Zhong and Lambowitz, 2003). The orientation bias for TeI4c was examined by cloning its DNA target site in opposite directions relative to the direction of plasmid replication (Figure 4.1). The recipient constructs were designed to contain the DNA target sequence in the sense (LEAD) and anti-sense (LAG) orientations.

The TeI4c intron is interrupted by the insertion of a ORF-less intron, TeI3c (Mohr et al., 2010). This set of nested introns is known as a twintron (Dai and Zimmerly, 2002a; Mohr et al., 2010). The TeI4c IEP is able to mobilize both the TeI4c intron (~2.6 x 10^{-2} mobility efficiency) and ORF-less TeI3c intron (~ 6.7% mobility efficiency) (Mohr et al., 2010). This portion of the study examines the ability of the TeI4c Δ502 IEP to promote retrohoming of the TeI3c and TeI4c introns through endonuclease-independent pathways. The mobility of the TeI4c Δ502 IEP was compared with the full-length TeI4c IEP. Donor constructs were designed that encoded either a TeI3c or TeI4c intron upstream of either a full-length or TeI4c Δ502 IEP.

The donor constructs were designed to encode either a TeI3c or TeI4c intron upstream of either the full-length or TeI4c Δ502 IEP (pACD42-3c4c Retarget, pADC42-3c4cΔ502 Retarget, pACD2x TeI4c4c, pACD2xTeI4c4c Δ502). The TeI3c intron donor construct, pACD42 3c4c Retarget, and its recipient, pBRRX +42, had been used previously for retargeting to optimize mobility (G. Mohr, unpublished data). Additionally, the mobility of the TeI4c IEP increases significantly when the TeI4b intron target site is used (Mohr et al., 2010). The recipient constructs were designed to contain the target sequence in both the sense (LEAD) and anti-sense (LAG) orientations (pBRRX+42 Lead/Lag, pBRR3T24b Lead/Lag). The donor and recipient vectors were co-transformed into *E. coli* HMS174(DE3) cells, induced at 48°C for 1 h, and plated onto LB (Luria-Bertani) plates containing tetracycline plus ampicillin or ampicillin alone. The mobility efficiency was calculated as the ratio of $(Tet^R + Amp^R)/Amp^R$ colonies after incubation at 37°C.

Figure 4.2 summarizes the mobility data for donor plasmids expressing full-length and TeI4c Δ502 protein. The full-length TeI4c IEP mobilized the TeI3c intron with a mobility efficiency of ~74% for both the leading and lagging strand constructs. The TeI4c Δ502 IEP mobilized the TeI3c intron with a mobility efficiency of ~13%, a 5-fold decrease from full-length TeI4c. The TeI4c 502 IEP with TeI3c intron also displayed a pronounced replication orientation bias (~8 fold) for using a nascent leading DNA strand as a primer during reverse transcription.

The full-length TeI4c protein with TeI4c intron had a mobility efficiency of ~30% for both the leading and lagging strand constructs. The TeI4c Δ502 IEP with TeI4c intron

had a mobility efficiency of ~1.5%, a 20-fold decrease from full-length TeI4c IEP. The TeI4c Δ502 IEP with TeI4c intron also displayed a replication orientation bias (~10 fold) for the nascent leading strand.

Together, the results above show that the TeI4c full-length protein, with either the TeI3c or TeI4c introns, does not have a strand preference for priming reverse transcription. However, the TeI4c Δ502 IEP has a leading strand preference with both introns. The lack of bias for the full-length TeI4c protein for both introns indicates that retrohoming occurs by reverse splicing into double-strand DNA and use of the endonuclease domain for bottom-strand cleavage to generate the primer for reverse transcription. The leading-strand bias for TeI4c Δ502 IEP for both introns indicate that in the absence of endonuclease cleavage retrohoming occurs preferentially by reverse splicing into double-stranded DNA and use of a nascent leading strand as a primer for reverse transcription.

4.2 Endonuclease activity of TeI4c

The *in vitro* endonuclease activity of the TeI4c IEP with the TeI3c and TeI4c intron was assayed by incubating ribonucleoprotein (RNP) particles with small (100 and 88 bp respectively) end-labeled DNA substrates. The lariat RNPs were prepared through protein-assisted splicing (full-length and TeI4c Δ502) of the transcribed intron precursor at 50°C for 1 h, followed by ultracentrifugation overnight. The ^{32}P end-labeled DNA substrate was made by designing long oligonucleotides, sense and anti-sense, that when annealed together contained the intron insertion site (ligated E1 and E2). The oligonucleotides (sense and anti-sense) were first ^{32}P end-labeled and then annealed with

their respective complementary unlabeled oligonucleotide to make the double-stranded DNA substrate. The endonuclease assay was performed by incubating the lariat RNP particle with the DNA substrate at 50°C for one hour. The products were phenol extracted, ethanol precipitated, and analyzed in a 6% polyacrylamide denaturing gel.

The DNA substrate utilized was end-labeled (5'-sense strand and 5'-antisense strand). Thus, the following products should be observed in the polyacrylamide gel: the complete reverse-spliced DNA, substrate, 5'-top and 5'-bottom (Figure 4.3 and 4.4). The 5'-top is the result of partial reverse-splicing of the RNP, and the 5'-bottom from bottom strand cleavage by the endonuclease domain. The full-length TeI4c protein, TeI3c and TeI4c intron RNPs, were active and reverse-spliced into the substrate DNA. However, for both the TeI3c and TeI4c intron endonuclease assays, a 5'-bottom product was not observed. This suggests that either TeI4c RNPs do not have endonuclease activity, the RNPs have endonuclease activity but the assay is not sensitive enough to detect less specific cleavage, or conditions were not optimal for the assay. The RNPs made with the TeI4cΔ502 IEP, and either TeI3c or TeI4c intron, were not active for reverse splicing and endonuclease bottom-strand cleavage *in vitro*.

4.3 Discussion

The TeI4c Δ502 IEP is able to retrohome through endonuclease-independent pathways. The TeI4c Δ502 protein has a pronounced replication orientation bias consistent with using the leading strand at the replication fork as a primer for reverse transcription. This preference arises from the intron RNA reverse splicing into a double-stranded DNA template, which positions the RNP to directly use the nascent leading-

strand at the replication fork as a primer for reverse transcription (San Filippo and Lambowitz, 2002). Reverse splicing into double-stranded DNA is biased against using the lagging strand primer, as the replication fork would first have to travel past the inserted RNP before using the lagging nascent strand as a primer. The inserted RNP might impede passage of the replication fork. Additionally, the replication fork might disrupt the inserted RNP, reducing the mobility efficiency of lagging strand retrohoming (San Filippo and Lambowitz, 2002).

The full-length TeI4c protein, unlike the TeI4c Δ502 protein, did not display a strand preference for using the leading or lagging strand at a replication fork. Thus, it seems likely that the full-length TeI4c protein is able to promote reverse splicing into double-stranded DNA and produce a primer through bottom-strand cleavage, which could then be used for reverse transcription. The group II intron-encoded protein from L1.LtrB, LtrA exhibits a similar leading strand preference when it lacks endonuclease activity. These results suggest that the TeI4c full-length protein uses its endonuclease domain to perform bottom-strand cleavage equally on both the leading and lagging strands.

RNPs containing the TeI4c IEP with either TeI3c and TeI4c intron RNAs could reverse splice into double-stranded DNA *in vitro*. However, the TeI4c Δ502 RNPs were inactive and unable to reverse splice into double-stranded DNA. Whether the IEP contains endonuclease activity is still unknown. Mobility assays suggest that the TeI4c protein uses its endonuclease domain during retrohoming to produce a primer used in reverse transcription. Thus, it is likely that the TeI4c IEP has endonuclease activity, but

the assay was not sensitive enough to detect less specific bottom-strand cleavage, perhaps because it is less specific and does not occur at a single site. It is also possible that the assay conditions were not optimal for bottom-strand cleavage.

4.4 Materials and Methods

Recombinant plasmids

The recipient vectors pBRR3B-LtrB (LEAD) and pBRR3A-LtrB (LAG) had been previously designed to contain the DNA target sites cloned in opposite orientations relative to the direction of plasmid direction (Zhong and Lambowitz, 2003). The intron target site recipient vectors (pBRRx+42, pBRR3T2-4b) were constructed by swapping the TeI3c and TeI4b target sequences for the LtrB target sequences in recipient vectors pBRR3B-LtrB (LEAD) and pBRR3A-LtrB (LAG). This was done through restriction enzyme digestion with AatII and SphI, followed by T4 DNA ligation.

The donor plasmid pACD2xTeI4c-4cΔ502 was constructed by digesting the plasmid pACD2xTeI4c-4c with NdeI and XhoI (Mohr et al., 2010). The TeI4cΔ502 ORF was PCR amplified by primers that append restriction sites for NdeI and XhoI. The PCR products were ligated into the digested pACD2xTeI4c-4c vector. The pADC42TeI3c4cΔ502 retargeted vector was made by digestion of the pACD2xTeI4c-4cΔ502 vector with NcoI and NdeI to remove TeI4cΔ502. The NcoI and NdeI fragment containing the TeI4c Δ502 ORF was then ligated into the NcoI and NdeI digested vector pADC42TeI3c4c (Mohr et al., 2010).

Intron mobility assays

Intron mobility assays were performed in *E. coli* HMS174(DE3) competent cells (Novagen, Madison, WI) as previously described (Mohr et al., 2010). The following concentrations of antibiotics were used in the assay: ampicillin, 100 μg/ml; chloramphenicol, 25 μg/ml; tetracycline, 25 μg/ml. Cells were co-transformed with the Cap^R donor and the Amp^R recipient plasmid and were grown in 5 ml of LB media containing ampicillin and chloramphenicol overnight at 37°C. A fraction (100 μl) of the overnight culture was used to inoculate 5 ml of fresh LB media containing appropriate antibiotics and grown for 1 h at 37°C. The cells were then induced for 1 h at 48°C with 1ml of LB containing a final IPTG concentration of 500 μM. The cells were then placed on ice, diluted with ice-cold LB media, and plated at different dilutions onto LB agar plates containing ampicillin and tetracycline or ampicillin alone. The plates were incubated overnight at 37°C, colonies were counted, and the mobility efficiency was calculated as the ratio of $(Tet^R + Amp^R)/(Amp^R)$ colonies.

Reconstitution of group II intron RNPs

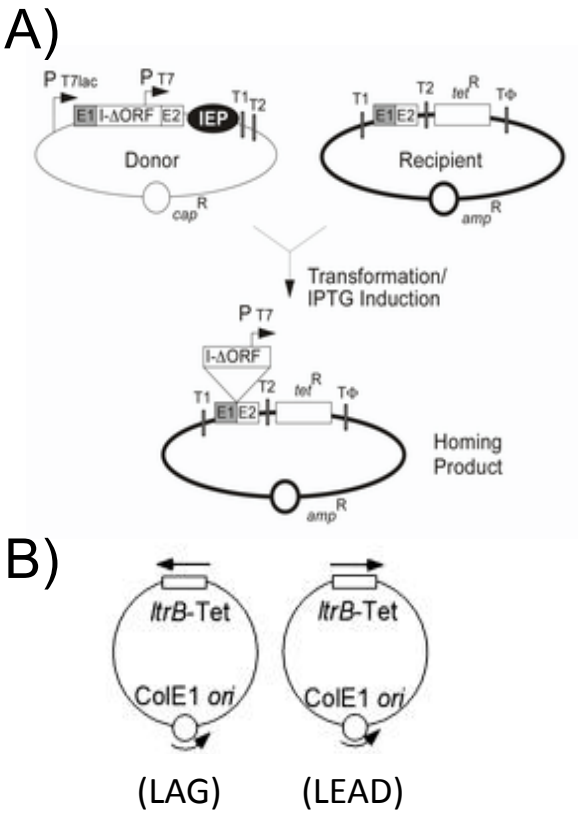
Lariat RNPs were made through protein-assisted splicing using purified TeI4c or TeI4c Δ502 proteins. The intron transcript (TeI3c and TeI4c) was transcribed overnight at 37°C. The RNA transcript (100 nM) was denatured by heating at 82°C for 2 min, and then cooling to 50°C. Buffer (10 mM Tris-HCl pH 7.5, 5 mM $MgCl_2$, 450 mM ammonium chloride) was added to the splicing reaction, followed by protein (200 nM). The protein and transcript were incubated at 50°C for 1 h, and spun in the ultracentrifuge (Beckman 70.1 Ti rotor) at 50 K at 4°C overnight. The lariat RNPs were resuspended in

endonuclease buffer (10 mM KCl, 10 mM $MgCl_2$, 50 mM Tris-HCl pH 7.5) and stored at -80°C.

Endonuclease assay

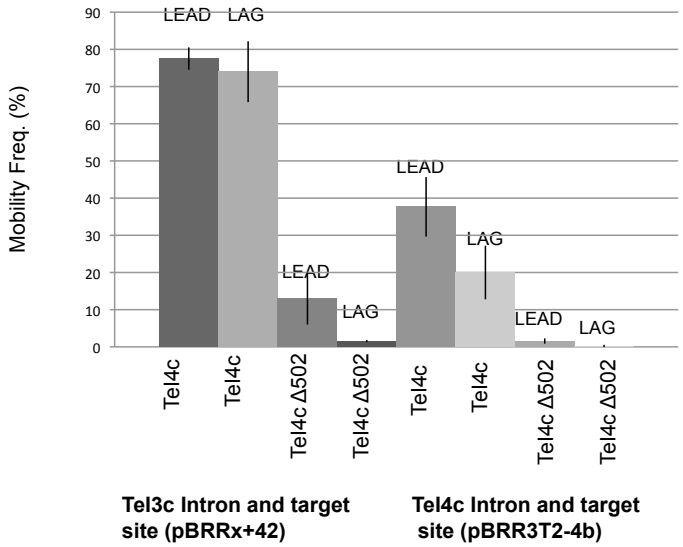
Complementary oligonucleotides were designed that contained the intron insertion site (ligated E1 and E2) of the TeI3c and TeI4c introns. Each oligonucleotide was end-labeled using T4 polynucleotide kinase (NEB) and annealed with its complementary unlabeled oligonucleotide to make a double-stranded DNA substrate. The lariat RNP (20 pmol) was incubated with the DNA substrate (1 pmol) in endonuclease buffer at 50°C for 1 h. The reaction was then phenol extracted, precipitated with ethanol, and analyzed on a 6% polyacrylamide denaturing gel.

Figure 4.1: Plasmid-based intron mobility assay



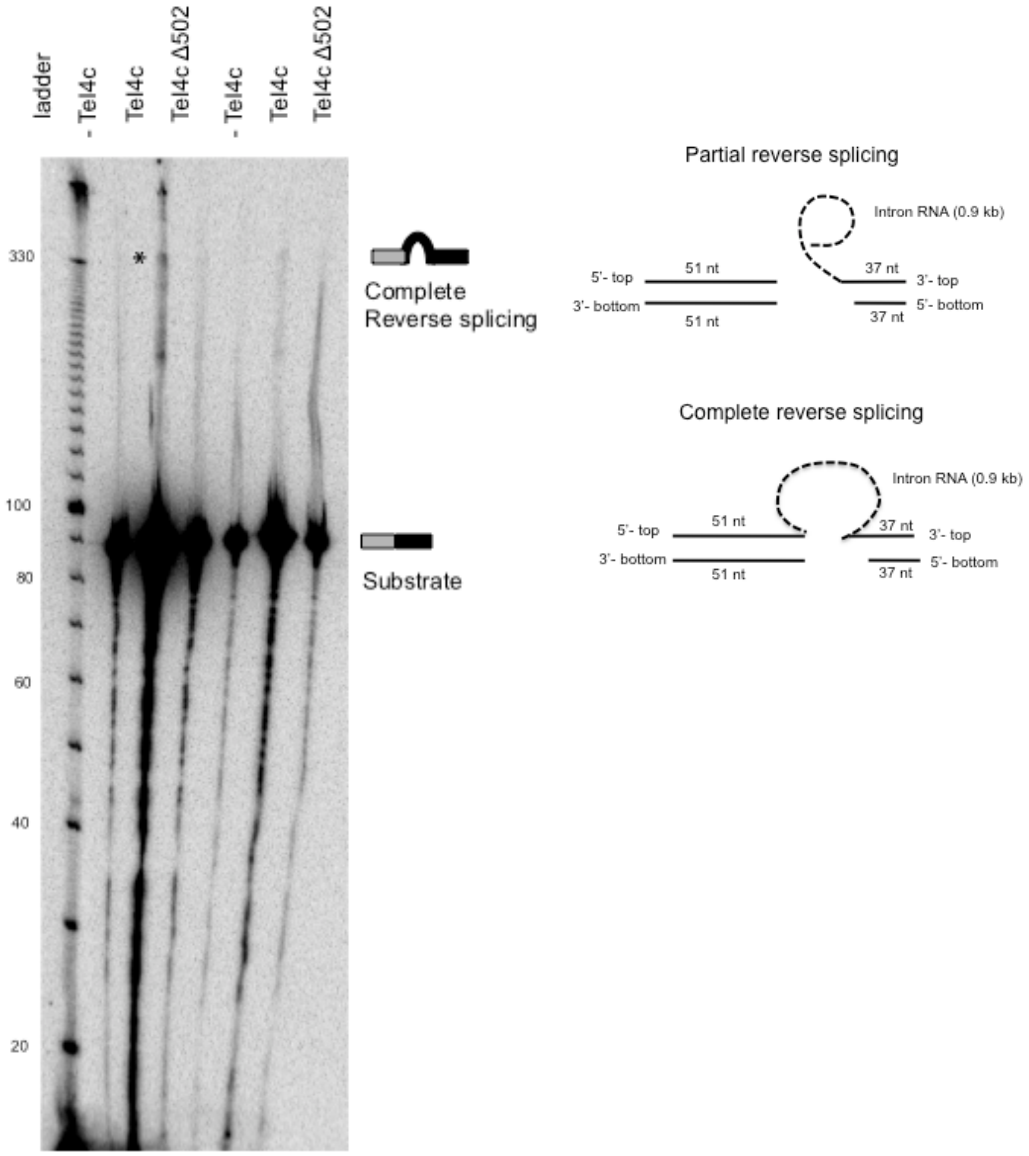
(A) The donor construct is a chloramphenicol resistant (Cap^R) plasmid containing the ORF-less intron and flanking exons. A T7 RNA polymerase promoter is located in DIV near the 3' end of the intron. The IEP is downstream of the intron. The recipient construct is an ampicillin resistant (Amp^R) plasmid containing the target site (ligated E1 and E2) cloned upstream of a promoter-less tetracycline (Tet^R) gene. When the intron, carrying the T7 promoter, integrates into the target site the tetracycline gene is activated. Adapted from Mohr et al. (2010).
(B) The recipient constructs were designed to contain the DNA target site and tet^R gene cloned in opposite orientations relative to the direction of plasmid replication. The plasmids were denoted leading (LEAD) or lagging (LAG) depending on whether the nascent leading or lagging strands could be used as a primer during reverse transcription. Adapted from Zhong and Lambowitz (2003).

Figure 4.2: Mobility assays on full-length and Δ502 TeI4c protein with TeI3c and TeI4c target sites



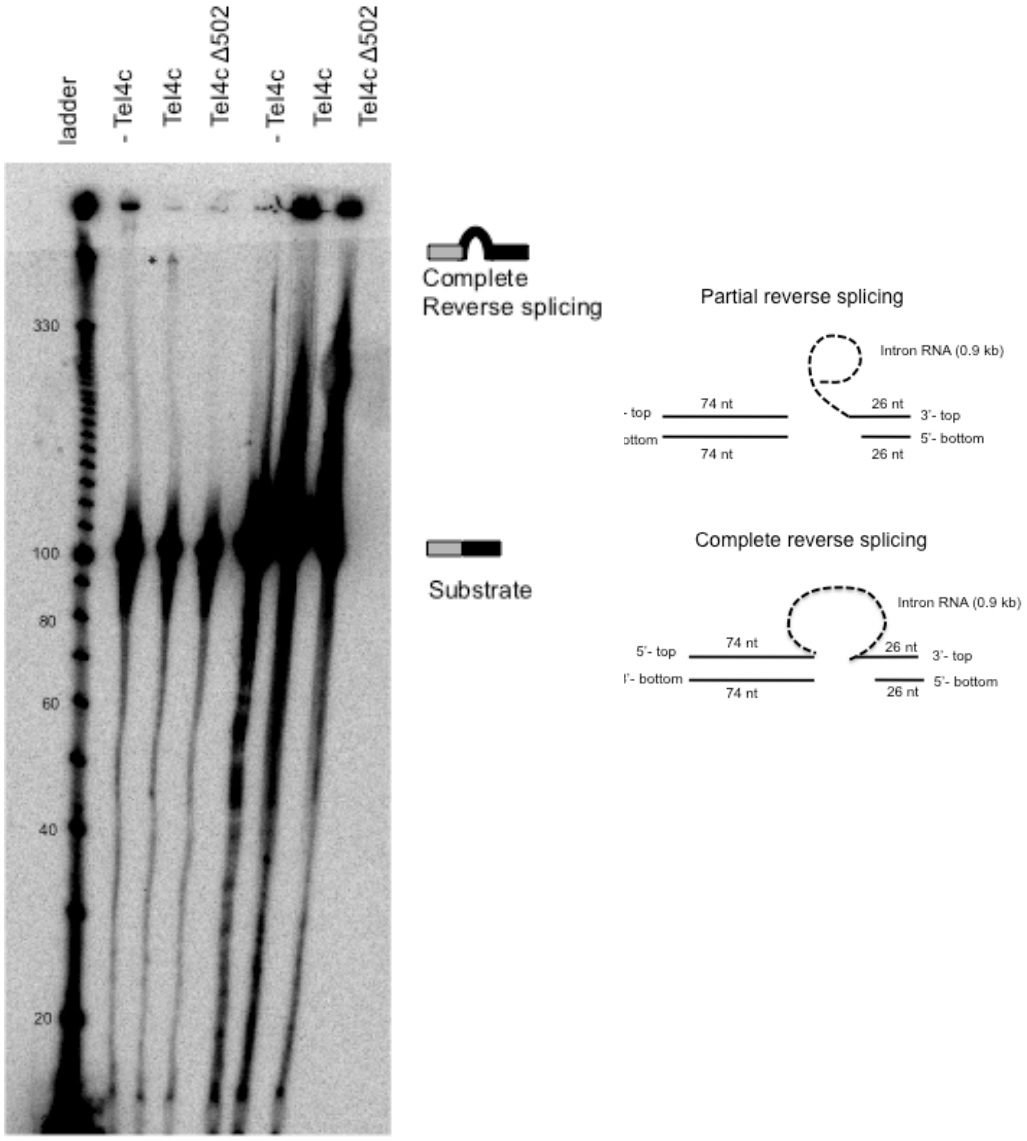
The mobility assays were performed with a two-plasmid mobility system, as described in Figure 4.1. The donor plasmids contained full-length TeI4c protein (pACD2x TeI4c4c, pACD42-3c4cRetarget) or truncated TeI4c Δ502 protein (pADC42-TeI3c4cΔ502Retarget, pACD2xTeI4c4c delta502). The recipient plasmids contained the TeI3c intron target site (pBRR3T2-4b Lead, and pBRR3T2-4b Lag) or TeI4c target site (pBRRx+42 Lead, pBRRx+42 Lag). The DNA target site was cloned in opposite orientations, Lead or Lag, relative to the direction of plasmid replication. The donor and recipient vectors were co-transformed in HMS174(DE3) competent cells, and induced with 500 μM IPTG at 48°C for 1 hour. The cells were spread onto plates containing tetracycline and ampicillin or ampicillin alone. The mobility efficiency was calculated as the ratio of $(Tet^R + Amp^P)/(Amp^R)$. The full-length and TeI4c Δ502 protein mobility assays were performed at the simultaneously. Each assay was repeated three times. The bars indicate standard deviation.

Figure 4.3: TeI4c IEP with TeI4c intron endonuclease assay



The endonuclease activity was assayed by incubating full-length or TeI4c Δ502 RNP particles with an 88-nucleotide DNA substrate at 50°C. The products were analyzed on a 6% polyacrylamide denaturing gel. The asterisk (*) indicates on the gel the full-length TeI4c protein complete reverse splicing product.

Figure 4.4: TeI4c IEP with TeI3c intron endonuclease assay



The endonuclease activity was assayed by incubating full-length or TeI4c Δ502 RNP particles with a 100-nucleotide DNA substrate at 50°C. The products were analyzed on a 6% polyacrylamide denaturing gel. The asterisk (*) indicates on the gel the full-length TeI4c protein complete reverse splicing product.

Chapter 5: Crystallization of a group II intron-encoded protein

Crystallization trials have been ongoing to determine the structure of the group II intron-encoded protein TeI4c. As discussed previously, the endonuclease domain is not required for reverse transcriptase activity, making the smaller MBP-RF-TeI4c Δ502 protein an ideal candidate for crystallization trials. Two techniques have been used in this study to increase the chances of crystal formation. The first technique is the use of a carrier protein, such as maltose-binding protein, to increase expression of the protein, aid in protein solubility and folding, and encourage the formation of a crystal lattice (Kapust and Waugh, 1999; Kobe et al., 1999; Smyth et al., 2003). The second technique aids crystallization by reducing the surface entropy (SER) of the maltose-binding protein (Derewenda, 2004). Since protein crystallization can be inhibited by the entropic behavior of large hydrophilic side chains on the protein surface, these side chains can be mutated to small nonpolar amino acids, reducing surface entropy and increasing the chance for protein crystallization (Avbelj and Fele, 1998). Five vectors containing surface entropy mutations in the maltose binding protein have been designed to enhance crystallization (Moon et al., 2010) (See Table 1.1). The pMalE-RF-TeI4cΔ502 vector was constructed with surface entropy mutations in the maltose-binding protein (Table 5.1).

5.1 TeI4cΔ502 crystallization trials

The pMalE-RF-TeI4cΔ502 expression plasmids and modified plasmids with the maltose-binding protein surface entropy mutations (Table 5.1) were used to express proteins in crystallization trials. The vectors were expressed in *E. coli* ArcticExpress

competent cells in LB medium and purified through an amlyose-affinity column and heperin-Sepharose column. The peak protein fractions were pooled and dialyzed overnight into buffer containing KCl, Tris-HCl pH 7.5, and glycerol. The protein was concentrated using a centrifugal protein filter (Millipore) at 4°C. The protein was then used to set-up 96-well sitting drop crystal trays at 22°C.

Several protein buffer conditions were tested to find the optimal conditions to set-down crystal trays. These conditions varied from 200 – 500 mM KCl, and 0 – 20% glycerol. The buffer condition, 20 mM Tris-HCl pH 7.5, 250 mM KCl, and 10% glycerol, kept the protein soluble and concentrated. The protein concentration were varied from 1 to 20 mg/ml and set-down in crystal trays. The optimal protein concentration to set-down trays was experimentally determined to be 10 mg/ml. Crystal trays were set-down utilizing commercially available 96-well screen blocks containing various salts and precipitants for high-throughput analysis of crystallization conditions (Crystal Screen HT, Index HT, PEG/ION HT, Grid Screen Salt, Salt Rx HT and PEG Rx HT) (Hampton). No crystallization screen tested resulted in TeI4c crystals. Other approaches will need to be used to explored to crystallize TeI4c.

5.2 Discussion

Although initial attempts to crystallize TeI4cΔ502 have been unsuccessful, other techniques could be utilized to form crystals. Full-length TeI4c or other TeI4c truncations, such as catalytically active TeI4cΔ484, could be used in crystallization trials. Additionally, a different carrier protein such as N utilization substance A (NusA), which maintains reverse transcriptase activity, could be used used in crystallization trials.

Surface-entropy mutations in TeI4c may also increase crystal contacts, allowing crystals to form. Surface entropy mutations can be predicted using online programs, such as SERp, that predict mutation candidates likely to enhance the protein's crystal contacts (Goldschmidt et al., 2007). Lastly, the addition of a DNA or RNA substrate may encourage crystallization.

5.3 Materials and Methods

Expression vectors

The pMal-RF-TeI4cΔ502 surface entropy reduction (SER) constructs were constructed through restriction enzyme digestion of the pMalE-RF-TeI4cΔ502 and pMBP-SER vector cassettes with BlpI and PstI (Moon et al., 2010). The BlpI/PstI fragment containing the TeI4cΔ502 ORF was then ligated into the BlpI and PstI MBP-SER digested vector, and the constructs were verified through DNA sequencing.

Protein expression and purification

pMalE-RF-TeI4cΔ502 and the maltose binding surface entropy mutants were expressed and purified as previously described for the pMalE-RF-TeI4cΔ502 protein with the following differences: the plasmids were transformed into ArcticExpress RIL (Agilent Technologies) competent cells, grown in LB medium at 30°C to an optical density (OD_{600}) of ~ 1.2-1.6, and expression was induced with 1 mM IPTG at 16°C for 24h.

Crystallization

TeI4c protein was dialyzed into various buffers containing 20 mM Tris pH 7.5, 200 mM- 500 mM KCl, and 0- 20% glycerol. The protein was concentrated from 1 – 20

mg/ml. Sitting drops were set-down at 22°C using commercially available 96-well crystallography screens containing various salts and precipitants (Crystal Screen HT, Index HT, PEG/ION HT, Grid Screen Salt, and PEG Rx HT) (Hampton). For each sitting drop, the reservoir contained 65 µl of crystallography screen solution. The pedestal contained a 1:1 or 1:2 mix of crystallization screen solution to protein.

Table 5.1: Expression vectors used in crystallization trials

Vector	SER mutations
pMal-RF-TeI4cΔ502	
pMal(A)-RF-TeI4cΔ502	D82A/K83A
pMal(B)-RF-TeI4cΔ502	E172A/N173A
pMal(C)-RF-TeI4cΔ502	D82A/K83A/K239A
pMal(D)-RF-TeI4cΔ502	E172A/N173A/K239A
pMal(E)-RF-TeI4cΔ502	D82A/K83A/E172A/N173A/K239A

The TeI4c Δ502 protein was expressed in vectors with and without various surface entropy reduction mutations (SER). The protein was expressed in *E. coli* ArcticExpress RIL (Agilent Technologies) competent cells, purified, and used in crystallization trials.

References

Avbelj, F., and Fele, L. (1998). Role of main-chain electrostatics, hydrophobic effect and side-chain conformational entropy in determining the secondary structure of proteins. Journal of molecular biology *279*, 665-684.

Bakhanashvili, M., and Hizi, A. (1993). The fidelity of the reverse transcriptases of human immunodeficiency viruses and murine leukemia virus, exhibited by the mispair extension frequencies, is sequence dependent and enzyme related. FEBS letters *319*, 201-205.

Blocker, F.J., Mohr, G., Conlan, L.H., Qi, L., Belfort, M., and Lambowitz, A.M. (2005). Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase. RNA *11*, 14-28.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., and Thompson, J.D. (2003). Multiple sequence alignment with the Clustal series of programs. Nucleic acids research *31*, 3497-3500.

Conlan, L.H., Stanger, M.J., Ichiyanagi, K., and Belfort, M. (2005). Localization, mobility and fidelity of retrotransposed Group II introns in rRNA genes. Nucleic acids research *33*, 5262-5270.

Costa, M., Michel, F., and Westhof, E. (2000). A three-dimensional perspective on exon binding by a group II self-splicing intron. The EMBO journal *19*, 5007-5018.

Dai, L., and Zimmerly, S. (2002a). Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. Nucleic acids research *30*, 1091-1102.

Dai, L., and Zimmerly, S. (2002b). The dispersal of five group II introns among natural populations of Escherichia coli. RNA *8*, 1294-1307.

Derewenda, Z.S. (2004). Rational protein crystallization by mutational surface engineering. Structure *12*, 529-535.

Dickson, L., Huang, H.R., Liu, L., Matsuura, M., Lambowitz, A.M., and Perlman, P.S. (2001). Retrotransposition of a yeast group II intron occurs by reverse splicing directly into ectopic DNA sites. Proceedings of the National Academy of Sciences of the United States of America *98*, 13207-13212.

Goldschmidt, L., Cooper, D.R., Derewenda, Z.S., and Eisenberg, D. (2007). Toward rational protein crystallization: A Web server for the design of crystallizable protein variants. Protein science : a publication of the Protein Society *16*, 1569-1576.

Gorbalenya, A.E. (1994). Self-splicing group I and group II introns encode homologous (putative) DNA endonucleases of a new family. Protein science : a publication of the Protein Society *3*, 1117-1120.

Guo, H., Karberg, M., Long, M., Jones, J.P., 3rd, Sullenger, B., and Lambowitz, A.M. (2000). Group II introns designed to insert into therapeutically relevant DNA target sites in human cells. Science *289*, 452-457.

Guo, H., Zimmerly, S., Perlman, P.S., and Lambowitz, A.M. (1997). Group II intron endonucleases use both RNA and protein subunits for recognition of specific sequences in double-stranded DNA. The EMBO journal *16*, 6835-6848.

Kapust, R.B., and Waugh, D.S. (1999). Escherichia coli maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. Protein science : a publication of the Protein Society *8*, 1668-1674.

Karberg, M., Guo, H., Zhong, J., Coon, R., Perutka, J., and Lambowitz, A.M. (2001). Group II introns as controllable gene targeting vectors for genetic manipulation of bacteria. Nature biotechnology *19*, 1162-1167.

Kobe, B., Center, R.J., Kemp, B.E., and Poumbourios, P. (1999). Crystal structure of human T cell leukemia virus type 1 gp21 ectodomain crystallized as a maltose-binding protein chimera reveals structural evolution of retroviral transmembrane proteins. Proceedings of the National Academy of Sciences of the United States of America *96*, 4319-4324.

Kohlstaedt, L.A., Wang, J., Friedman, J.M., Rice, P.A., and Steitz, T.A. (1992). Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. Science *256*, 1783-1790.

Lambowitz, A.M., and Zimmerly, S. (2004). Mobile group II introns. Annual review of genetics *38*, 1-35.

Lambowitz, A.M., and Zimmerly, S. (2010). Group II Introns: Mobile Ribozymes that Invade DNA. Cold Spring Harbor perspectives in biology.

Malik, H.S., Burke, W.D., and Eickbush, T.H. (1999). The age and evolution of non-LTR retrotransposable elements. Molecular biology and evolution *16*, 793-805.

Martinez-Abarca, F., and Toro, N. (2000). Group II introns in the bacterial world. Molecular microbiology *38*, 917-926.

Matsuura, M., Noah, J.W., and Lambowitz, A.M. (2001). Mechanism of maturase-promoted group II intron splicing. The EMBO journal *20*, 7259-7270.

Matsuura, M., Saldanha, R., Ma, H., Wank, H., Yang, J., Mohr, G., Cavanagh, S., Dunny, G.M., Belfort, M., and Lambowitz, A.M. (1997). A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: biochemical demonstration of maturase activity and insertion of new genetic information within the intron. Genes & development *11*, 2910-2924.

Michel, F., and Ferat, J.L. (1995). Structure and activities of group II introns. Annual review of biochemistry *64*, 435-461.

Michel, F., and Lang, B.F. (1985). Mitochondrial class II introns encode proteins related to the reverse transcriptases of retroviruses. Nature *316*, 641-643.

Mohr, G., Ghanem, E., and Lambowitz, A.M. (2010). Mechanisms used for genomic proliferation by thermophilic group II introns. PLoS biology *8*, e1000391.

Mohr, G., Smith, D., Belfort, M., and Lambowitz, A.M. (2000). Rules for DNA target-site recognition by a lactococcal group II intron enable retargeting of the intron to specific DNA sequences. Genes & development *14*, 559-573.

Moon, A.F., Mueller, G.A., Zhong, X., and Pedersen, L.C. (2010). A synergistic approach to protein crystallization: combination of a fixed-arm carrier with surface entropy reduction. Protein science : a publication of the Protein Society *19*, 901-913.

Munoz-Adelantado, E., San Filippo, J., Martinez-Abarca, F., Garcia-Rodriguez, F.M., Lambowitz, A.M., and Toro, N. (2003). Mobility of the Sinorhizobium meliloti group II intron RmInt1 occurs by reverse splicing into DNA, but requires an unknown reverse transcriptase priming mechanism. Journal of molecular biology *327*, 931-943.

Nakamura, Y., Kaneko, T., Sato, S., Ikeuchi, M., Katoh, H., Sasamoto, S., Watanabe, A., Iriguchi, M., Kawashima, K., Kimura, T., *et al*. (2002). Complete genome structure of the thermophilic cyanobacterium Thermosynechococcus elongatus BP-1. DNA research : an international journal for rapid publication of reports on genes and genomes *9*, 123-130.

Nallamsetty, S., and Waugh, D.S. (2006). Solubility-enhancing proteins MBP and NusA play a passive role in the folding of their fusion partners. Protein expression and purification *45*, 175-182.

Peebles, C.L., Perlman, P.S., Mecklenburg, K.L., Petrillo, M.L., Tabor, J.H., Jarrell, K.A., and Cheng, H.L. (1986). A self-splicing RNA excises an intron lariat. Cell *44*, 213-223.

Qin, P.Z., and Pyle, A.M. (1998). The architectural organization and mechanistic function of group II intron structural elements. Current opinion in structural biology *8*, 301-308.

Saldanha, R., Chen, B., Wank, H., Matsuura, M., Edwards, J., and Lambowitz, A.M. (1999). RNA and protein catalysis in group II intron splicing and mobility reactions using purified components. Biochemistry *38*, 9069-9083.

San Filippo, J., and Lambowitz, A.M. (2002). Characterization of the C-terminal DNA-binding/DNA endonuclease region of a group II intron-encoded protein. Journal of molecular biology *324*, 933-951.

Sarafianos, S.G., Marchand, B., Das, K., Himmel, D.M., Parniak, M.A., Hughes, S.H., and Arnold, E. (2009). Structure and function of HIV-1 reverse transcriptase: molecular mechanisms of polymerization and inhibition. Journal of molecular biology *385*, 693-713.

Schmelzer, C., and Schweyen, R.J. (1986). Self-splicing of group II introns in vitro: mapping of the branch point and mutational inhibition of lariat formation. Cell *46*, 557-565.

Shub, D.A., Goodrich-Blair, H., and Eddy, S.R. (1994). Amino acid sequence motif of group I intron endonucleases is conserved in open reading frames of group II introns. Trends in biochemical sciences *19*, 402-404.

Simon, D.M., Kelchner, S.A., and Zimmerly, S. (2009). A broadscale phylogenetic analysis of group II intron RNAs and intron-encoded reverse transcriptases. Molecular biology and evolution *26*, 2795-2808.

Singh, N.N., and Lambowitz, A.M. (2001). Interaction of a group II intron ribonucleoprotein endonuclease with its DNA target site investigated by DNA footprinting and modification interference. Journal of molecular biology *309*, 361-386.

Smith, D., Zhong, J., Matsuura, M., Lambowitz, A.M., and Belfort, M. (2005). Recruitment of host functions suggests a repair pathway for late steps in group II intron retrohoming. Genes & development *19*, 2477-2487.

Smyth, D.R., Mrozkiewicz, M.K., McGrath, W.J., Listwan, P., and Kobe, B. (2003). Crystal structures of fusion proteins with large-affinity tags. Protein science : a publication of the Protein Society *12*, 1313-1322.

Toro, N., Molina-Sanchez, M.D., and Fernandez-Lopez, M. (2002). Identification and characterization of bacterial class E group II introns. Gene *299*, 245-250.

Vellore, J., Moretz, S.E., and Lampson, B.C. (2004). A group II intron-type open reading frame from the thermophile Bacillus (Geobacillus) stearothermophilus encodes a heat-stable reverse transcriptase. Applied and environmental microbiology *70*, 7140-7147.

Wang, W., and Malcolm, B.A. (1999). Two-stage PCR protocol allowing introduction of multiple mutations, deletions and insertions using QuikChange Site-Directed Mutagenesis. BioTechniques *26*, 680-682.

Zhong, J., and Lambowitz, A.M. (2003). Group II intron mobility using nascent strands at DNA replication forks to prime reverse transcription. The EMBO journal *22*, 4555-4565.

Zhuang, F., Karberg, M., Perutka, J., and Lambowitz, A.M. (2009). EcI5, a group IIB intron with high retrohoming frequency: DNA target site recognition and use in gene targeting. RNA *15*, 432-449.

Zimmerly, S., Guo, H., Eskes, R., Yang, J., Perlman, P.S., and Lambowitz, A.M. (1995). A group II intron RNA is a catalytic component of a DNA endonuclease involved in intron mobility. Cell *83*, 529-538.

Zimmerly, S., Hausner, G., and Wu, X. (2001). Phylogenetic relationships among group II intron ORFs. Nucleic acids research *29*, 1238-1250.